Search by Sight:

Google™ goggles labs

Speaker: David Petrou

Alessandro Bissacco, Anand Pillai, Andrew Harp, Andrew Hogue, Andrew Rabinovich, Anthony Sciola, Casey Ho, Chuck Rosenberg, David Petrou, Fernando Brucher, Gabe Taubman, Hartmut Neven, Hartwig Adam, Henry Rowley, Jiayong Zhang, Johannes Steffens, John Flynn, Laura Garcia-Barrio, Lijia Jin, Matt Bridges, Matt Casey, Max Braun, Mihai Badoiu, Rafael Spring, Sergey Ioffe, Shailesh Nalawadi, Ulrich Buddemeier, Xiaotao Duan, Yuan Li

Themes

The What of Google Goggles

- Search by sight / Search by pointing

The How of Goggles

- What can we recognize and how well?
- What does not work?
- The process of Goggles

Digression on Augmented Reality

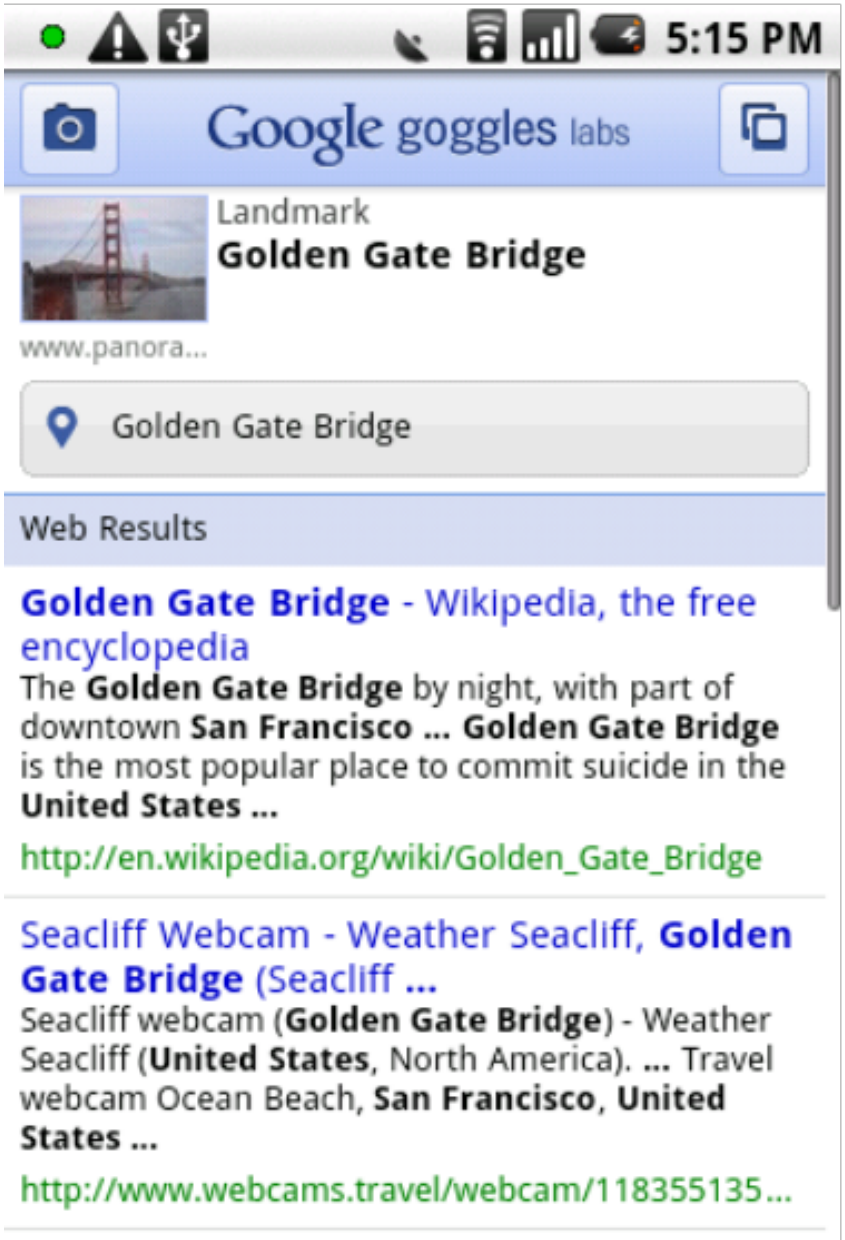
Performance

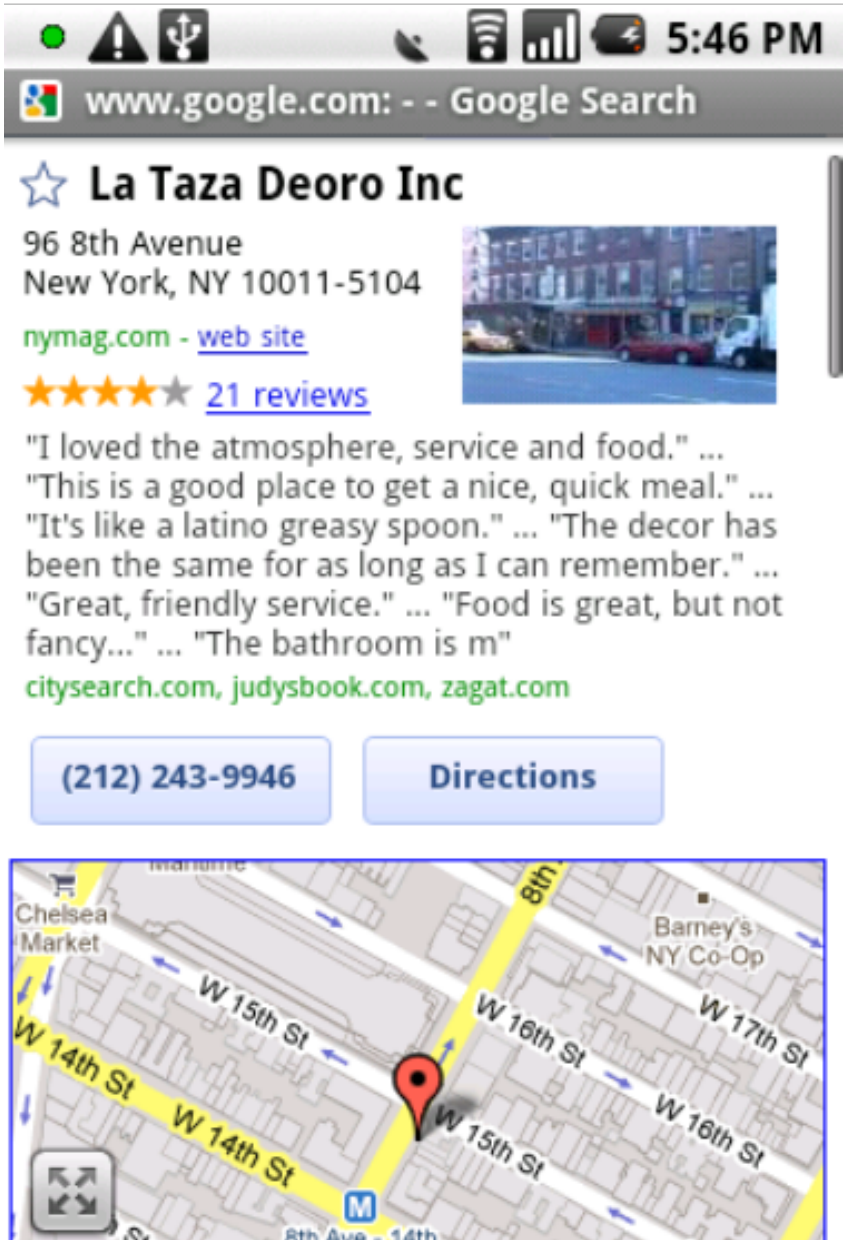
Where's Goggles going?

What is Google Goggles?

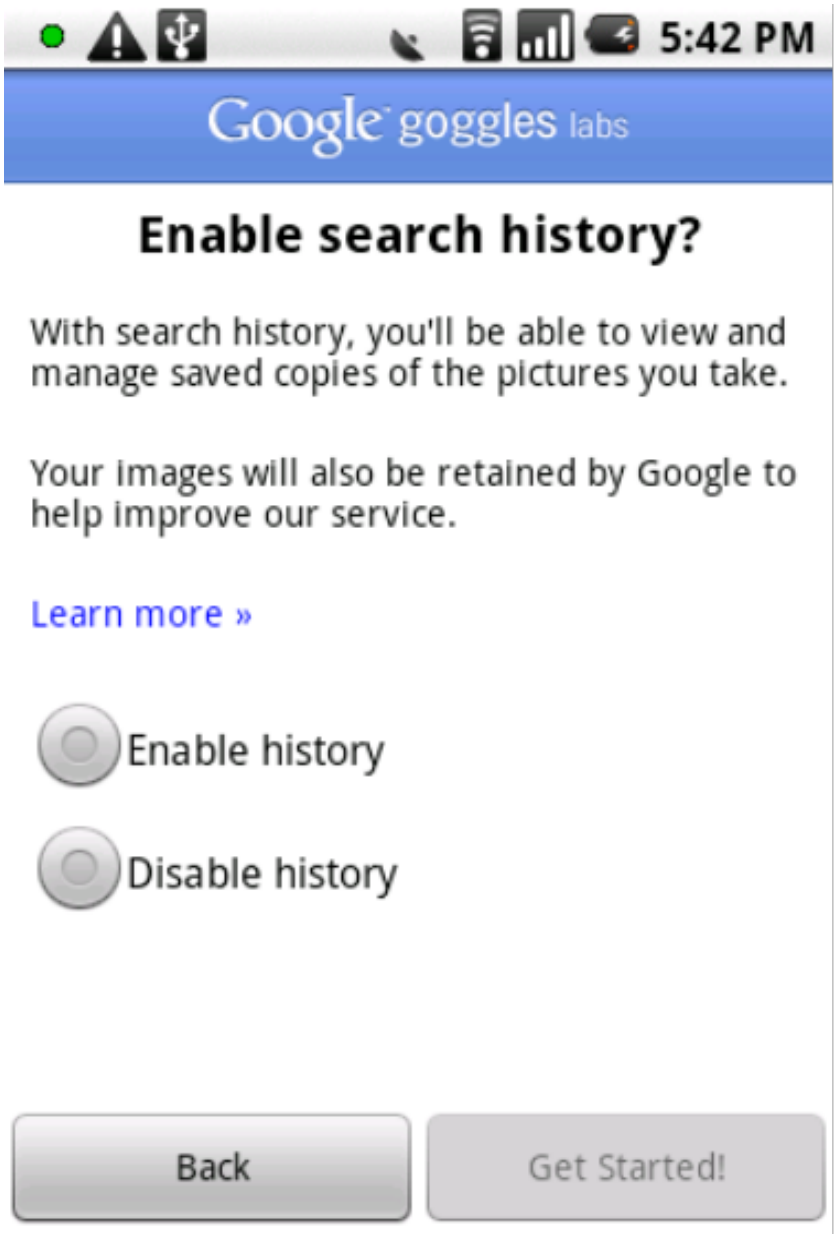
Google Goggles is a mobile visual search application currently available for Android mobile phones that lets a user submit a search query by taking a picture.

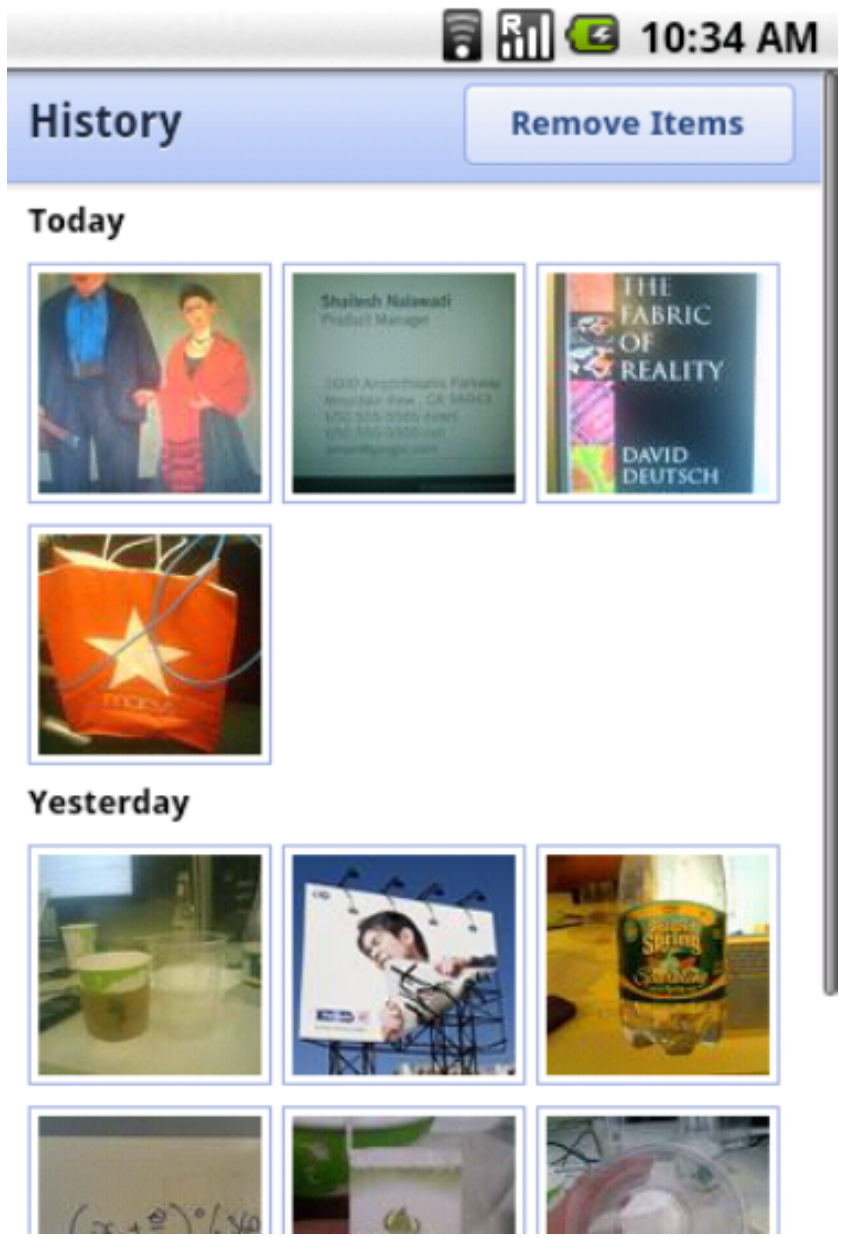
What's in front of you?





Visual search history





Use cases - Demos

I know what this is…bring it into my phone and tell me more

- OCR > canonical article, products, restaurants, ...
- Photo taking is faster than typing
- Barcodes problematic for advertising, so let's do more!

I don't know what this is…tell me

- Landmarks, long-tail objects, foreign language
- Impossible to type!

Unintended

- Trivia night

You are a superhero

Design Principles

Universal search

- Moment of truth for computer vision
- Painfully aware of all the things we can not recognize well
- Much harder than serving a particular vertical
- Plenty of opportunity for false positives

Results need to be specific

- Instance not class recognition

Put best foot forward, show best few results

- Dealing with ambiguity at several levels

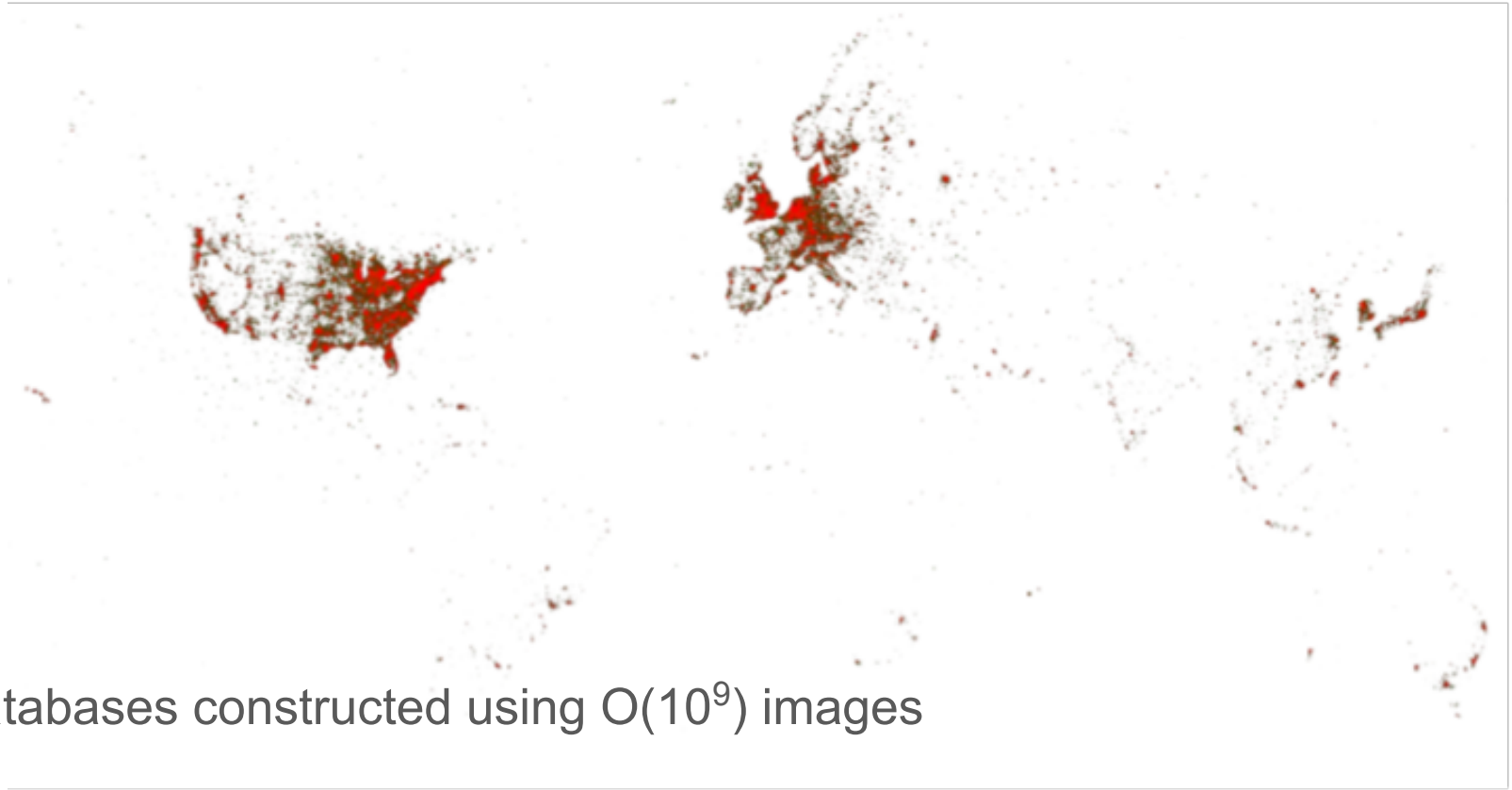
To the degree possible, do not force the user to select modes

Need the cloud for the corpora

Need image processing (GPS not sufficient)

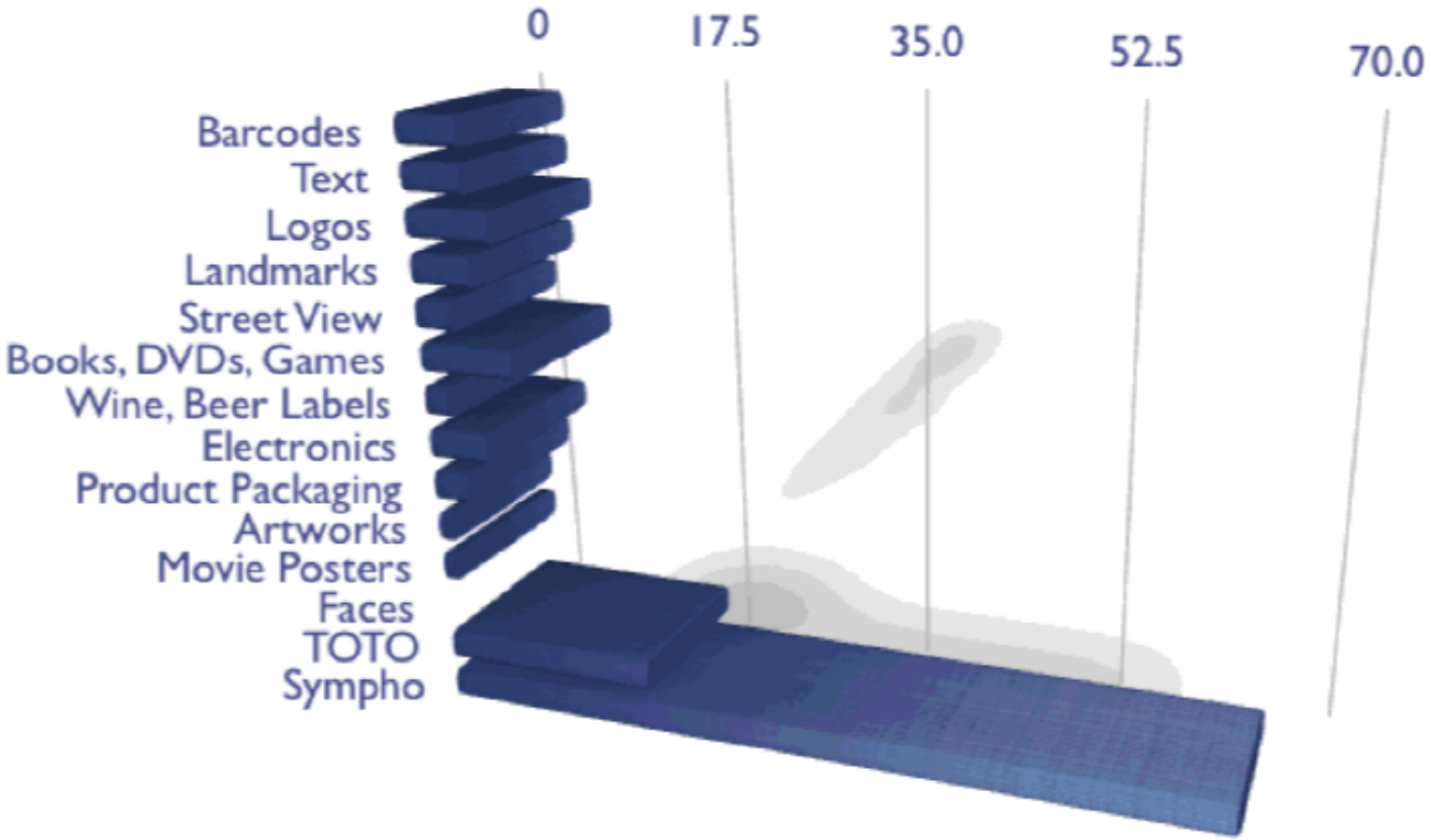
- Distinguishes Goggles from many Augmented Reality apps

Some statistics



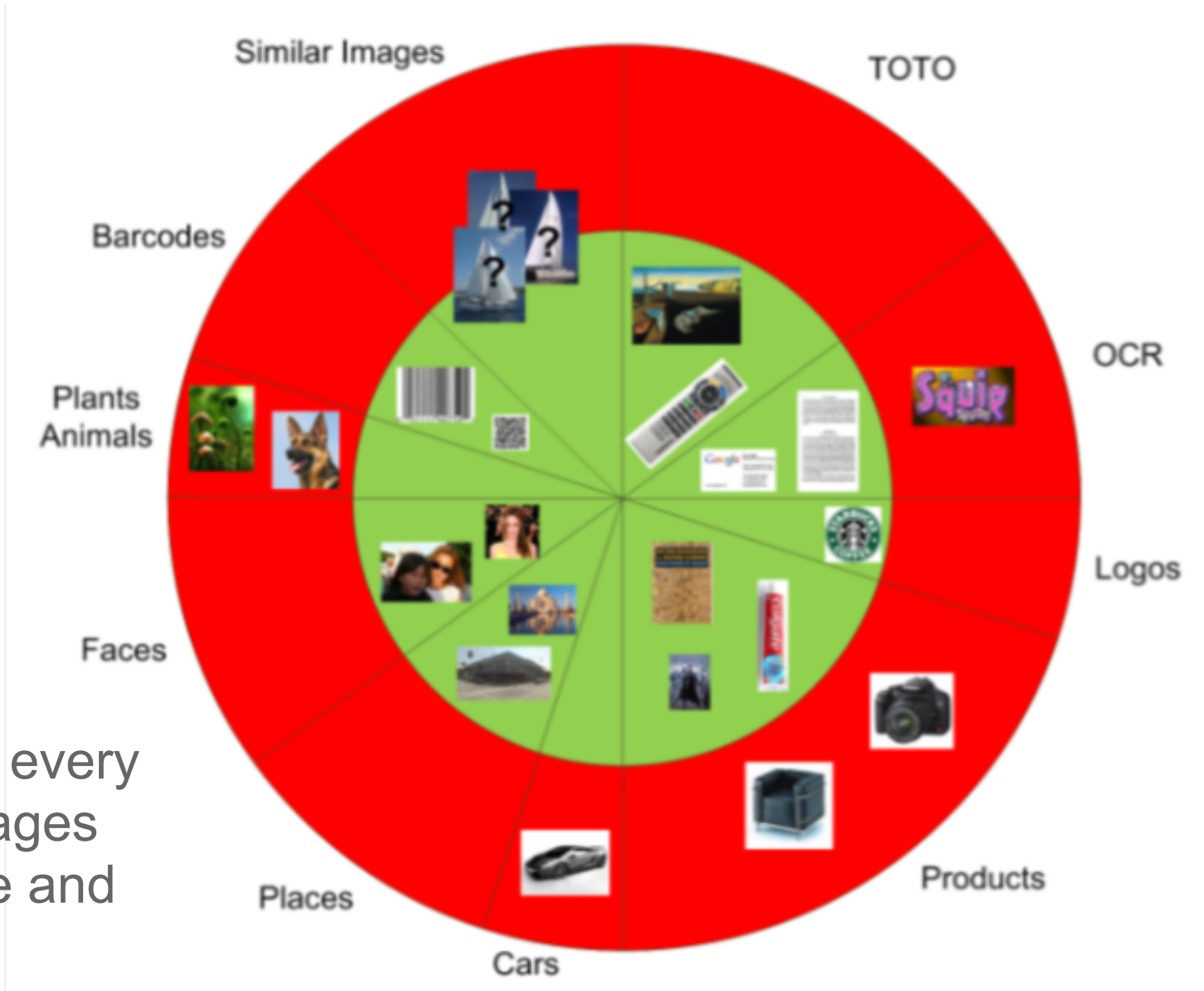
- Databases constructed using $O(10^9)$ images
- For about 33% of the queries we return a specific result
- In internal pre-launch testing >25% of queries included faces.
  Not supported until appropriate privacy models have been established!
- A lot of use in Hawaii ;-)

Recognition disciplines currently supported



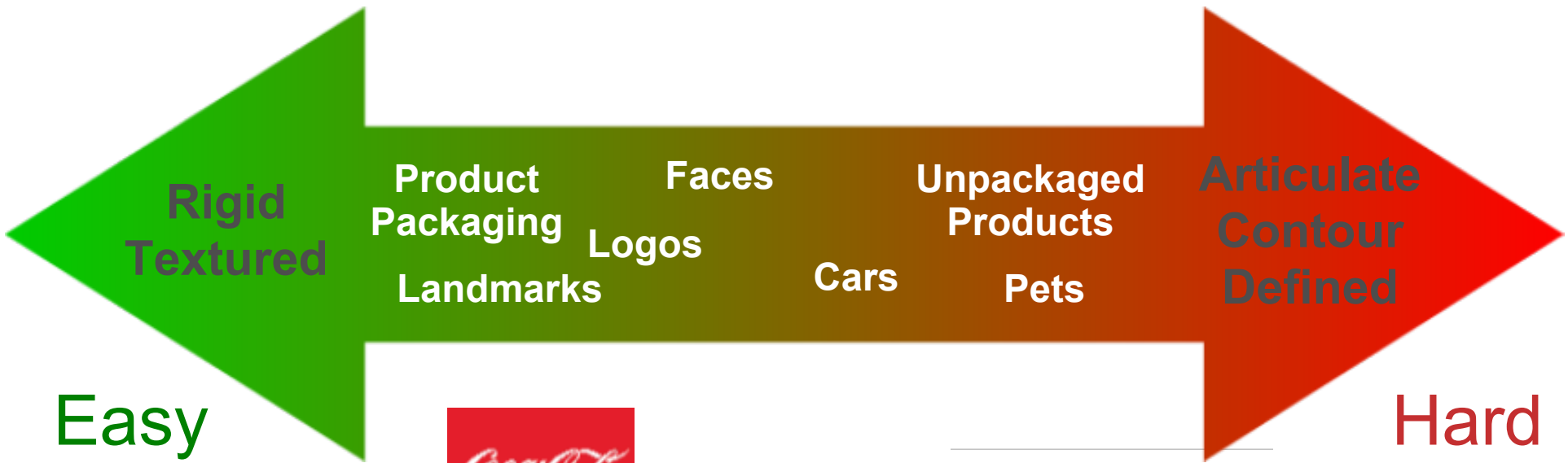
Multi-tunnel vision

Recognition disciplines that work and do not work



- Show a result for every query: similar images
- Learn from usage and ratings

Summary of the state of art (for product managers)

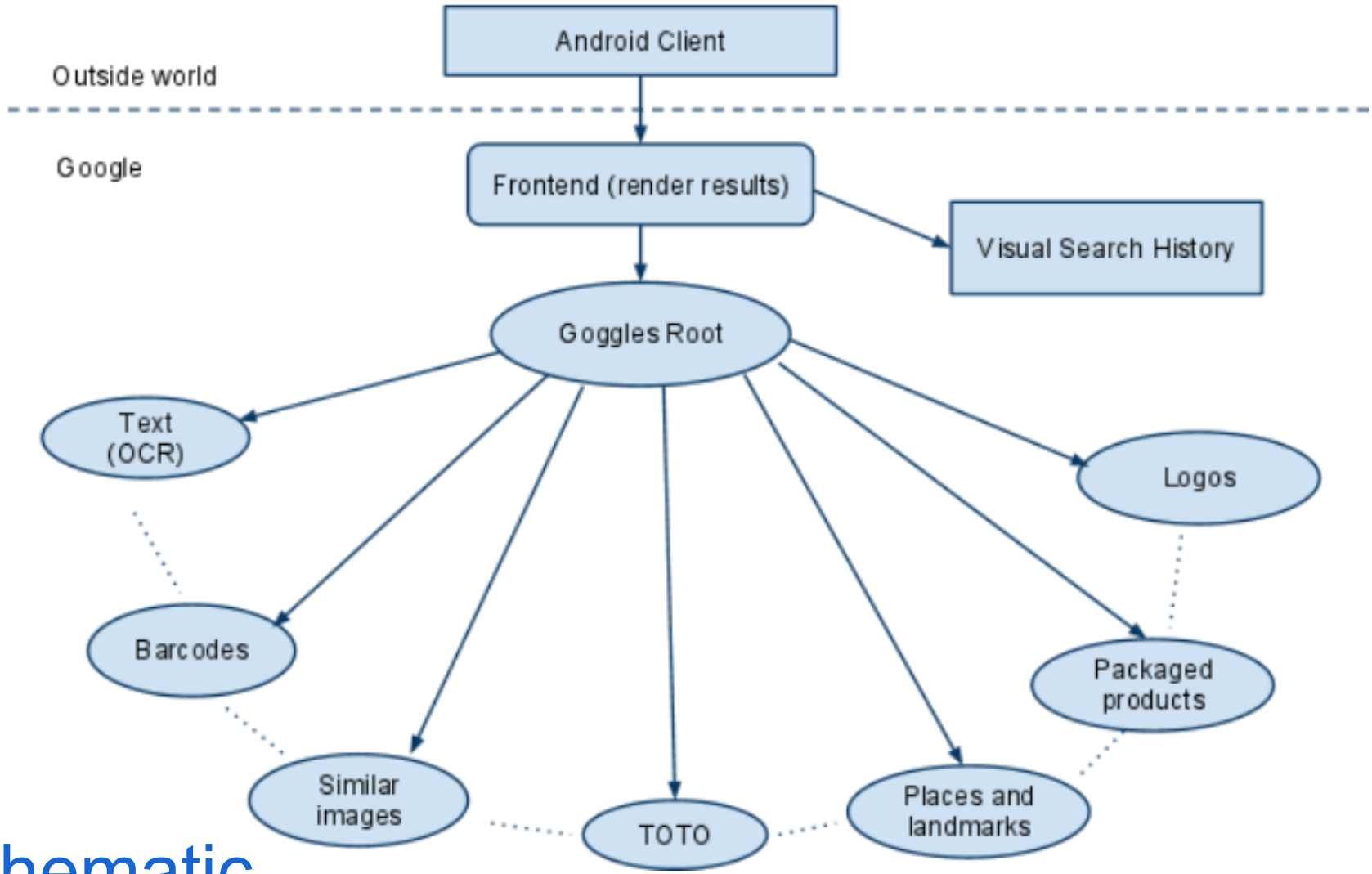


Easy

Hard

Recognition disciplines supported

Life of a Goggles query



Highly schematic
Highly parallel

The role of the Goggles Root

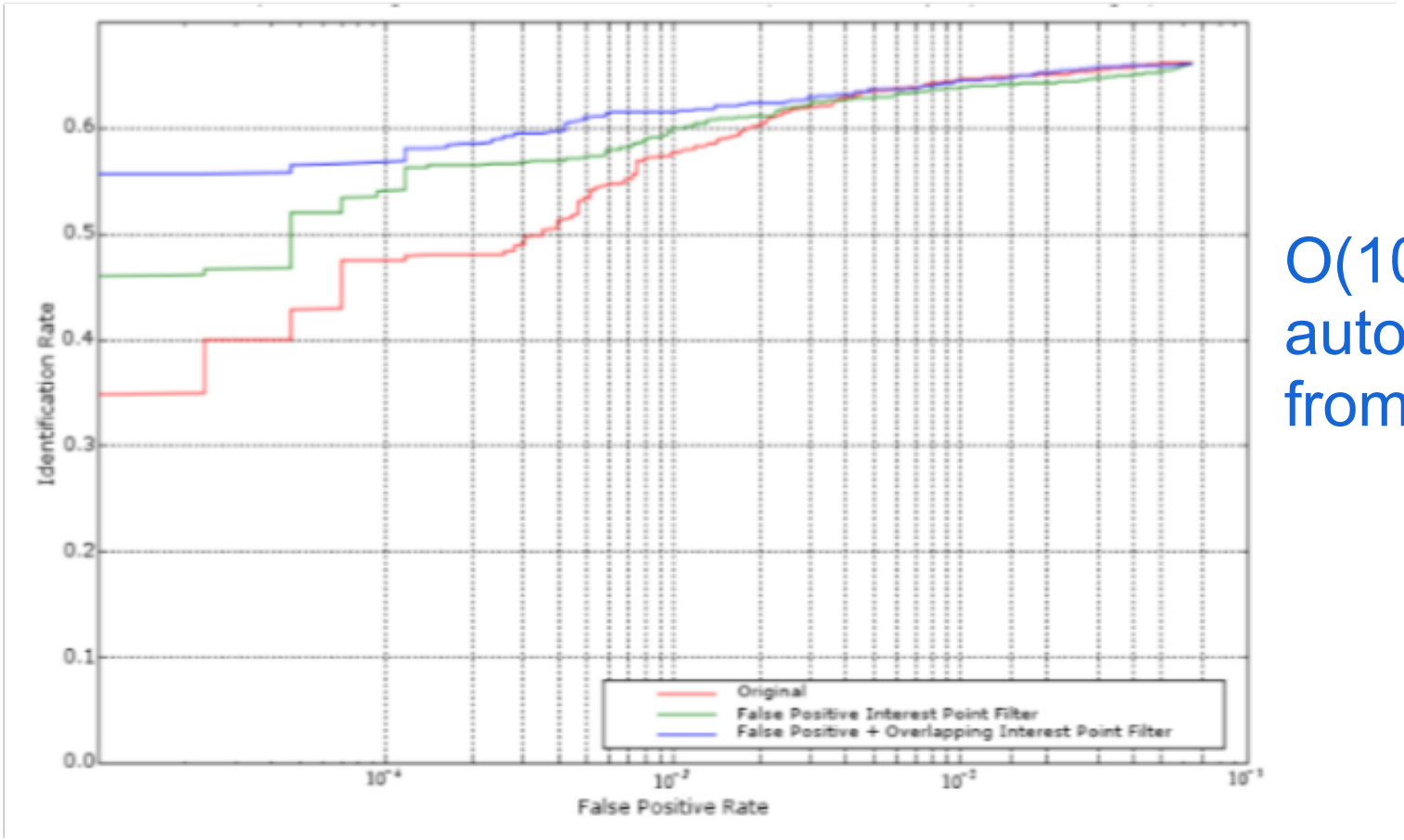
A picture is worth at least a thousand words
- How do we pick the best three?

Consider an image with many disparate objects
- Can we discern which object the user cares about?
- Use popularity of objects?
- Boost an object if OCR fires with text like its name?
- Use UI tricks? Led to the Goggles 1.2 UI

Consider one object that cause many backends to fire
- How can we combine the votes of many backends?

Textured object recognition

- Textured defined as sufficient number of interest points
- Supports many recognition channels
- CONGAS engine
- Can be driven to very low false positive rates



O(10^5) Place models automatically mined from photo collections

TOTO: **T**he **O**ther **T**extured **O**bjects

- Often QVGA resolution suffices
- Takes 0.2 sec per image
- $O(10^7)$ images from Image Search
  Discerning text labels: can feel like magic

Demo

Text recognition

Alternative input methodology

Add to contacts, fill in text fields, etc.

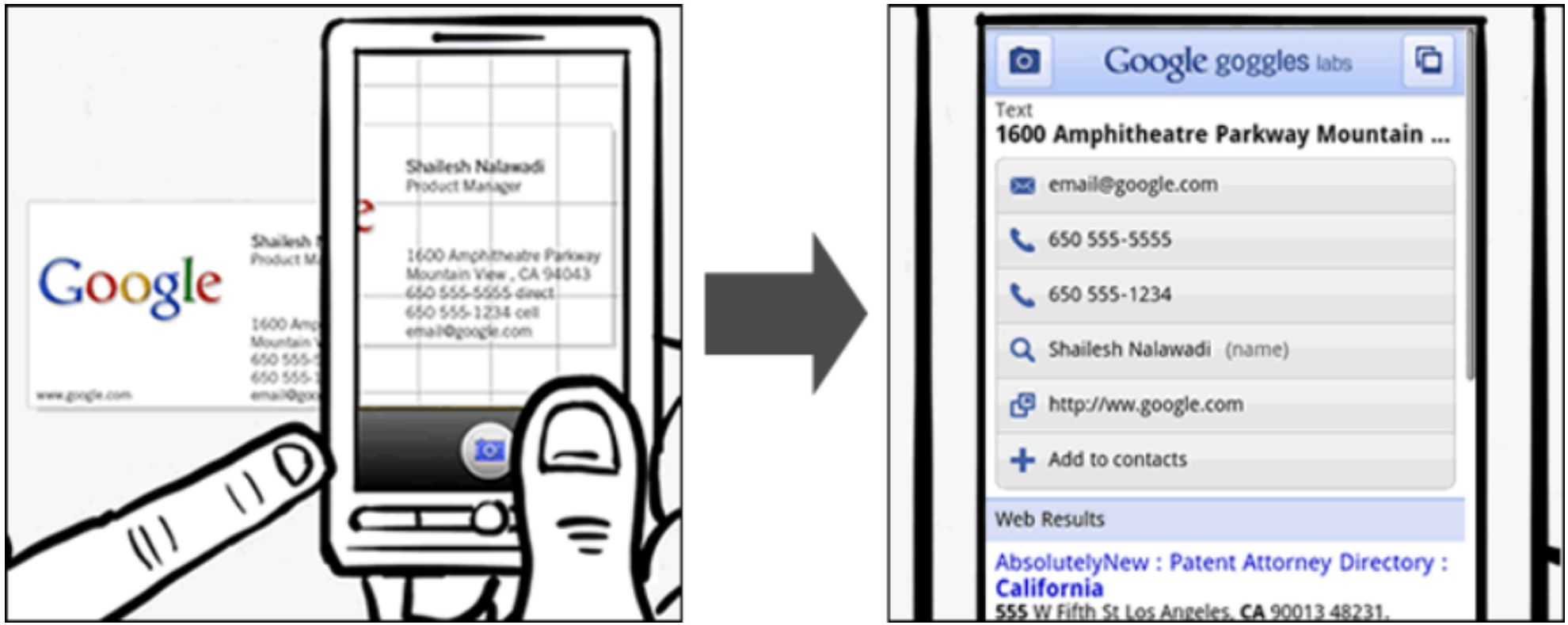
Translation

Foreign menus, travel signs, etc.

Result Support

Subtitle often identifies specific product

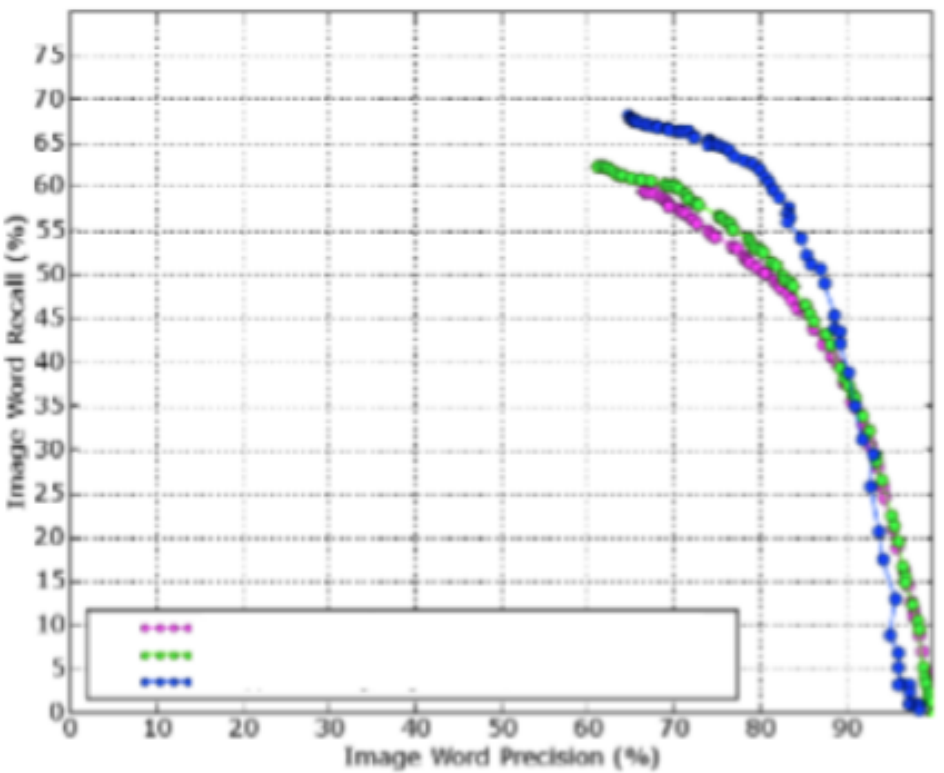
Multiple text uses
Entity extraction



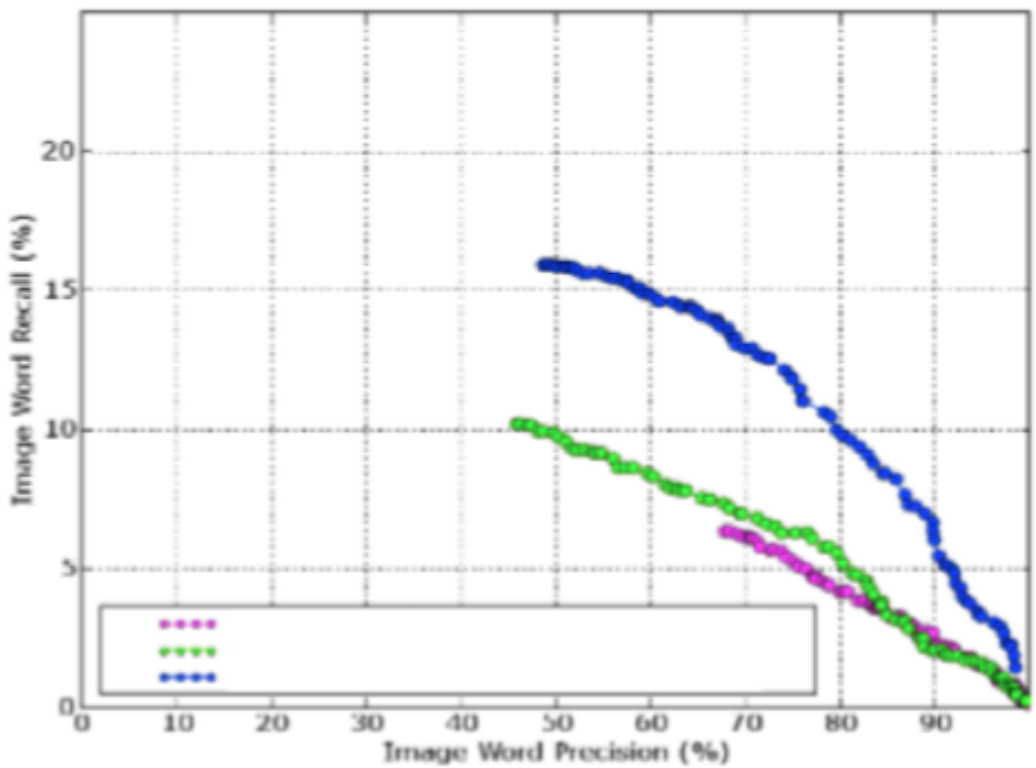
Multiple uses for text
Translation

Word level recall/precision of OCR

- Word level recall below 20%
- Needs VGA resolution
- Takes 3 sec per image

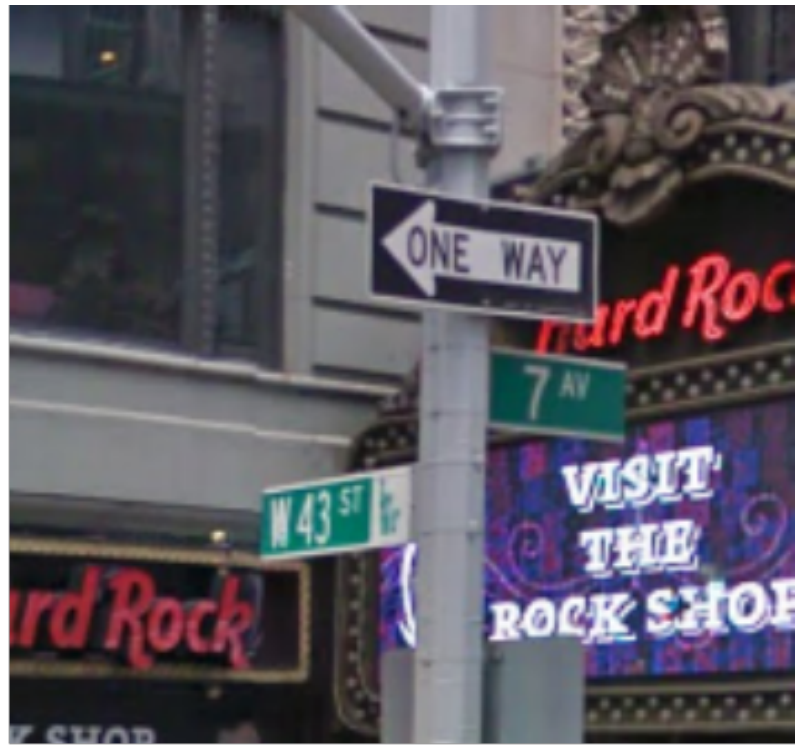


Dense



Sparse

Text recognition challenges



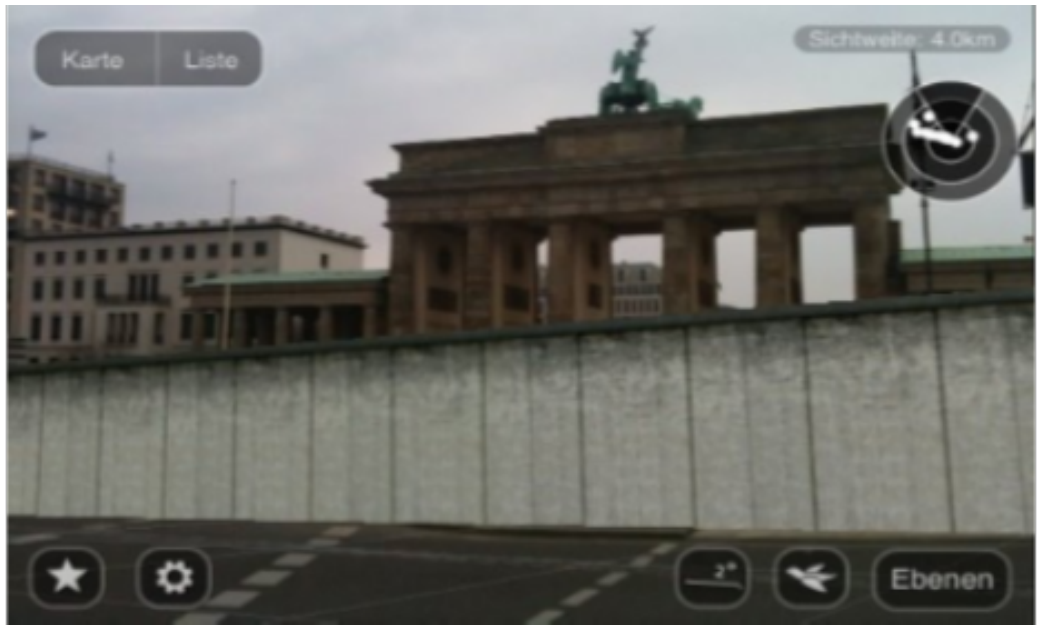
Curved text, handwritten, non-frontal, reflections, ...

The Process of Goggles

- Lots of different people contributing lots of different ideas to a really hard problem: universal visual search
- Problem comes up of how to combine these ideas
  - How to return relevant results, disambiguate user intent?
  - Achieved something that's more than the sum of its parts.
- Integration Science
- Balance between capability of technology and bits of creativity

Digression into Augmented Reality

- Examples:
  - Navigation information
  - User generated updates and themes: Layar, Tonchidot



Sananga, eye drop entheogen used by the Yawanawa tribe

Enables hunters to better see monkeys

Enables hunters to better see monkeys

Definition and context

Augmented Reality

View of a physical real-world
environment augmented by computer-generated imagery
R. Azuma 1997, P. Milgram and A. F. Kishino 1994

Generalization

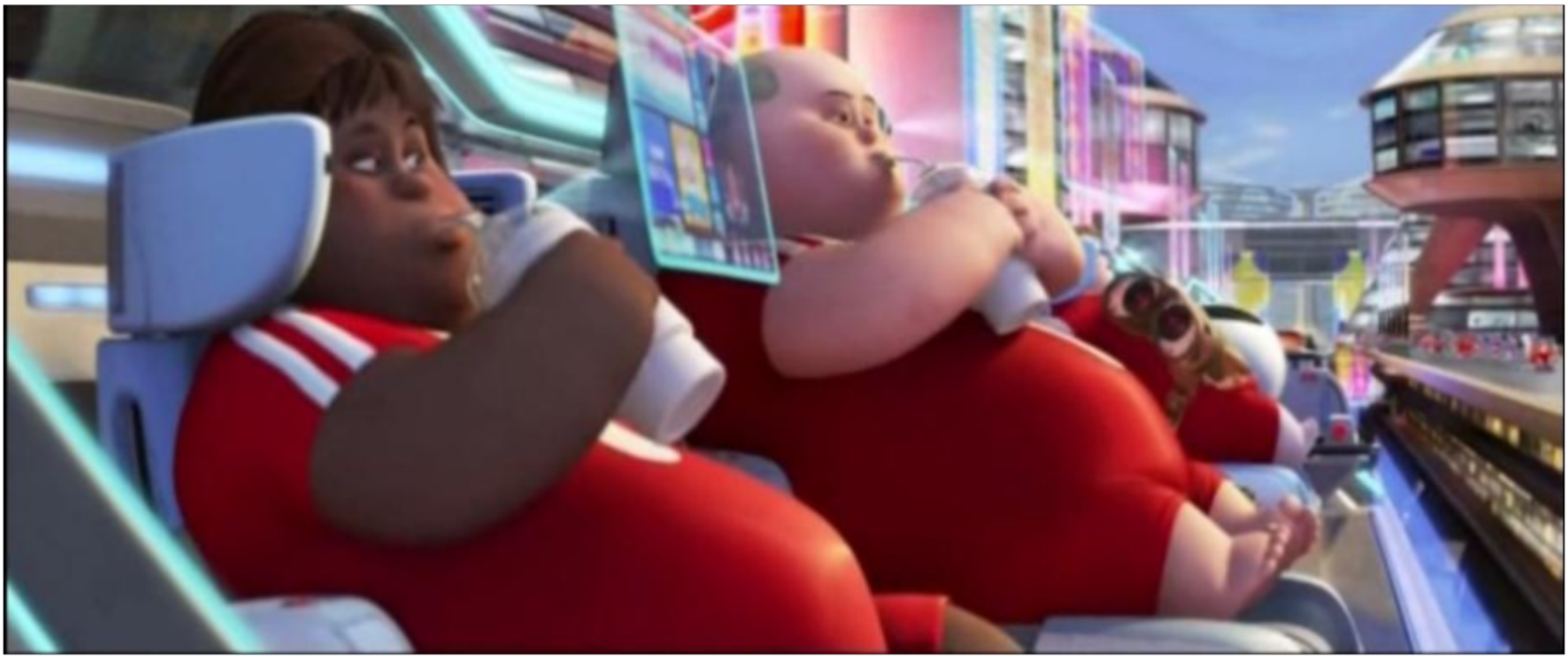
Presentation of information as a function of the environmental context

Not: Visual Search

Using an image as a search query

Not: Wearing a head mounted display

If this is our future, then maybe AR Is not all that important



Augmented Reality requires a physical reality one is interested in

Kaiser Family Foundation:
Today, 8-18 year-olds devote an average of **7 hours and 38 minutes** (7:38) to using entertainment media across a typical day (more than 53 hours a week).

And because they spend so much of that time 'media multitasking' (using more than one medium at a time), they actually manage to pack a total of **10 hours and 45 minutes** (10:45) worth of media content into those 7½ hours.

Augmented Reality requires a physical reality one is interested in

- Prerequisite is a novel object in my environment or the availability of novel information about an object nearby.
- Hence AR may not meet the bar of daily engagement.
- Same holds for visual search.

100:1 ratio between *internally* and *externally* triggered searches

- Less opportunity for externally triggered searches
  - Internal searches can be tucked away
  - External searches need immediate action
- If performed it feels magical and is often of high utility

- Most valuable when there is no voice or text substitute:
  - Faces (who is this colleague next to me?)
  - Disclaimer: Face recognition will only be offered once acceptable privacy models have been established
  - Restaurant in Tokyo for person who does not speak Japanese

- High convenience
  - Barcode
  - Add business card to contacts
  - Text for translation

Sources of Augmentation
From *Memory*

Marrying Goggles recognition with AR

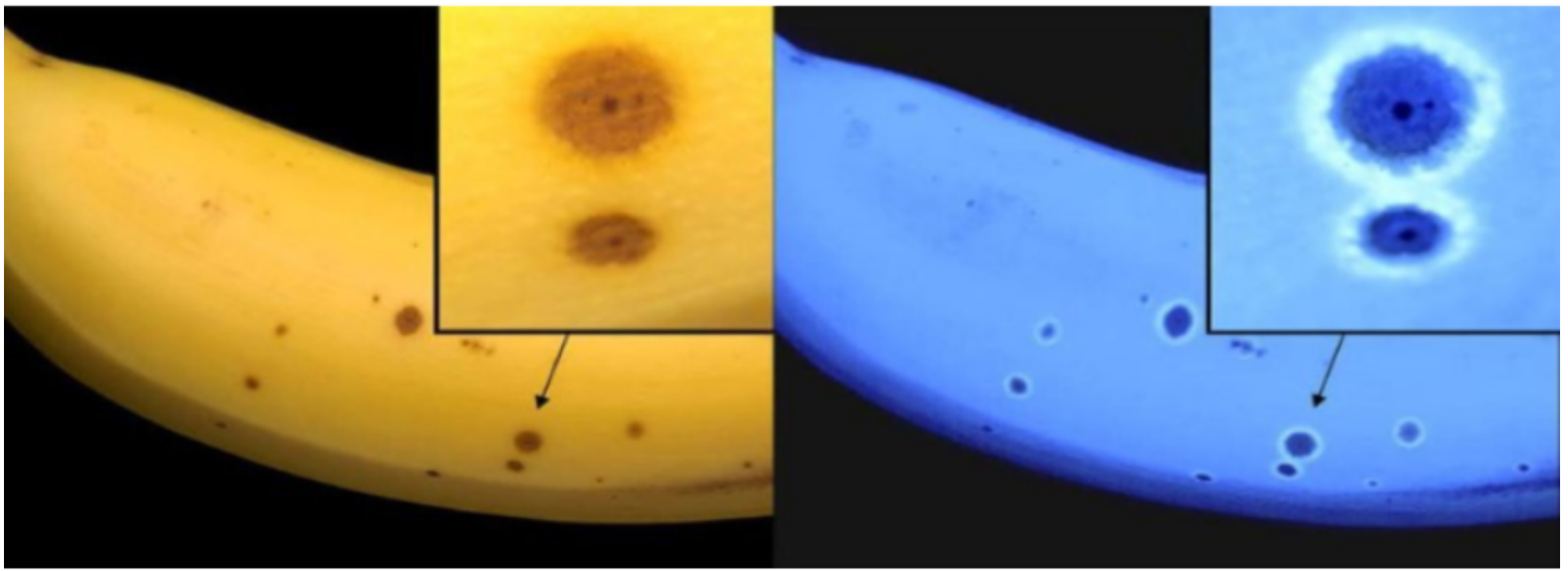
- AppSphere recognition
- Realtime interactive combination of virtual and physical, registered in 3D

Demo

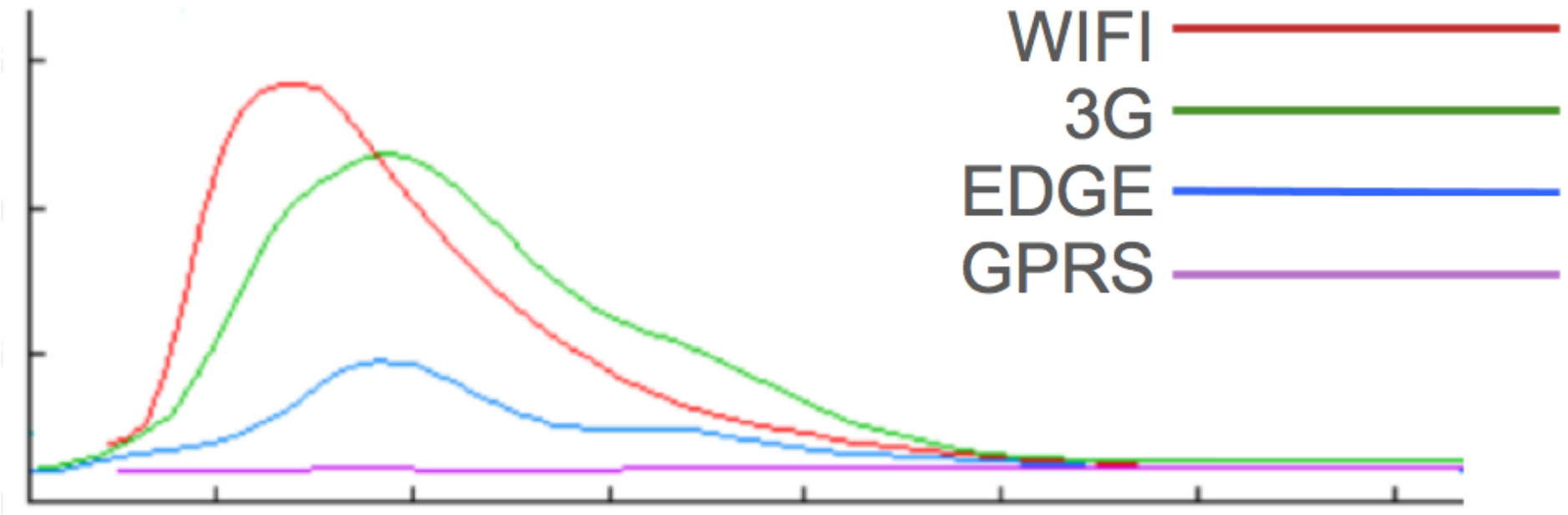
Sources of Augmentation From *Sensors*

Multispectral camera

- Infrared: find your pet
- Ultraviolet: Determine quality of food



Query delay by network type



time between queries

- The faster Goggles responds, the more usage we have
- It’s good to respond fast even (especially) when we can't give the right answer
  - The user will resend the query if the shot was unlucky

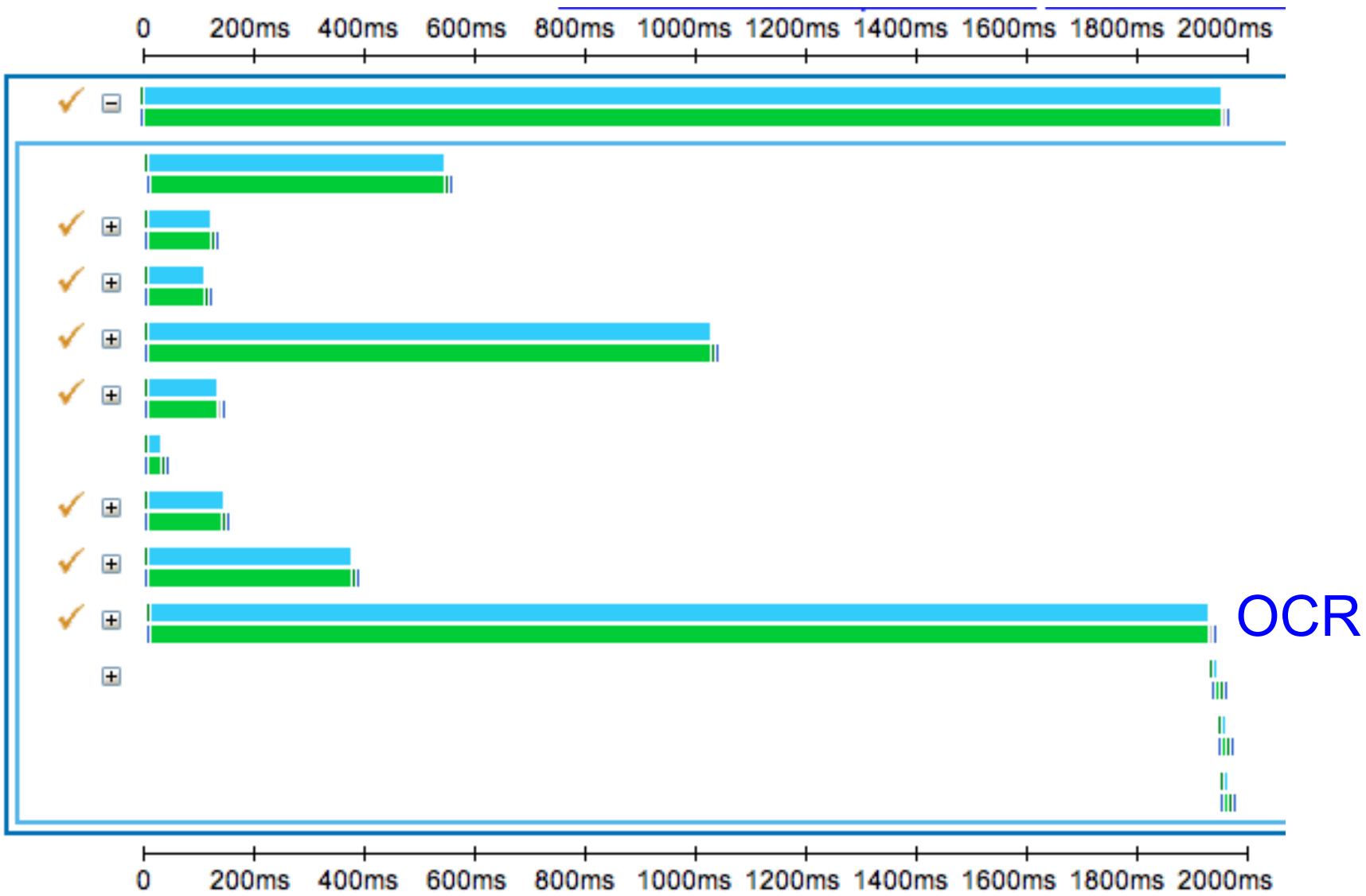
What determines latency?

- Client delays
  - Camera focusing: 1.5 s
  - Picture acquisition: 2 s
  - Image (re) encoding
- Network delay (upstream bandwidth): Variance!
- Highly heterogenous image recognition backends
- Rendering results

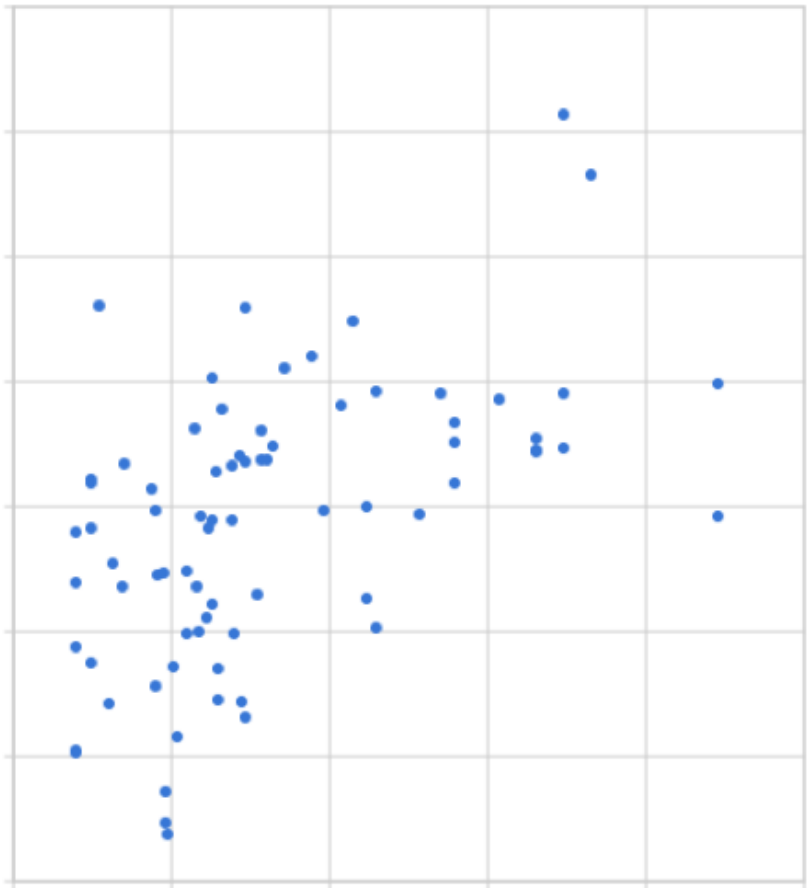
Problem: Goals at odds

- Universal visual search
  - Barcode and place recognition very different
- High quality visual search
  - OCR can use all the time it can get
- Low latency
  - Support browsing, fail-fast, user satisfaction

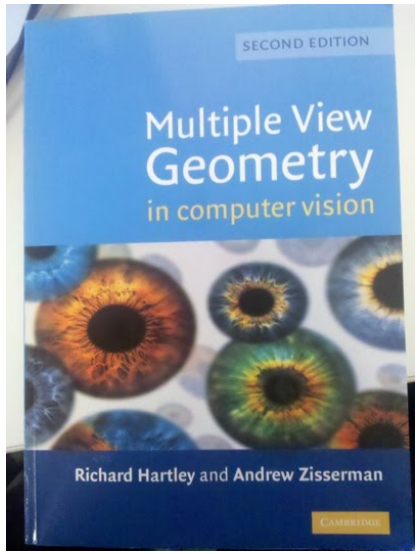
Server heterogeneity



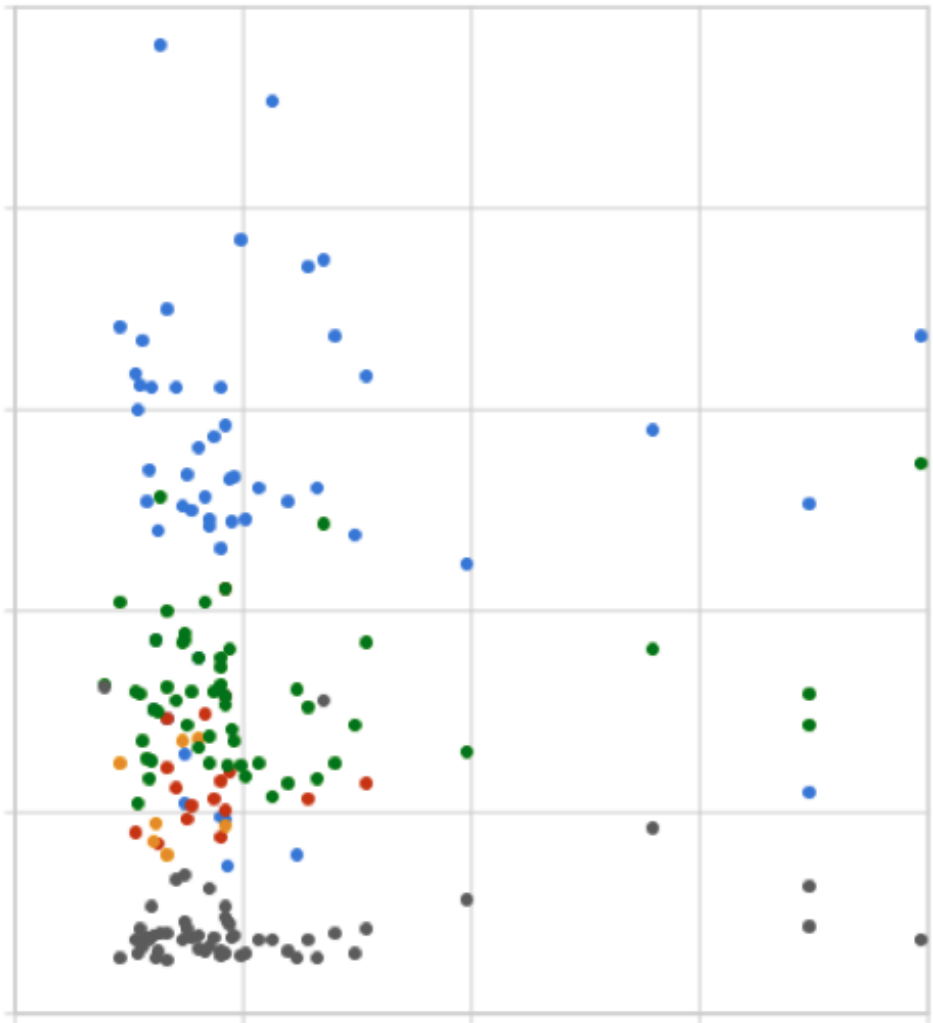
Book cover, waiting for all results



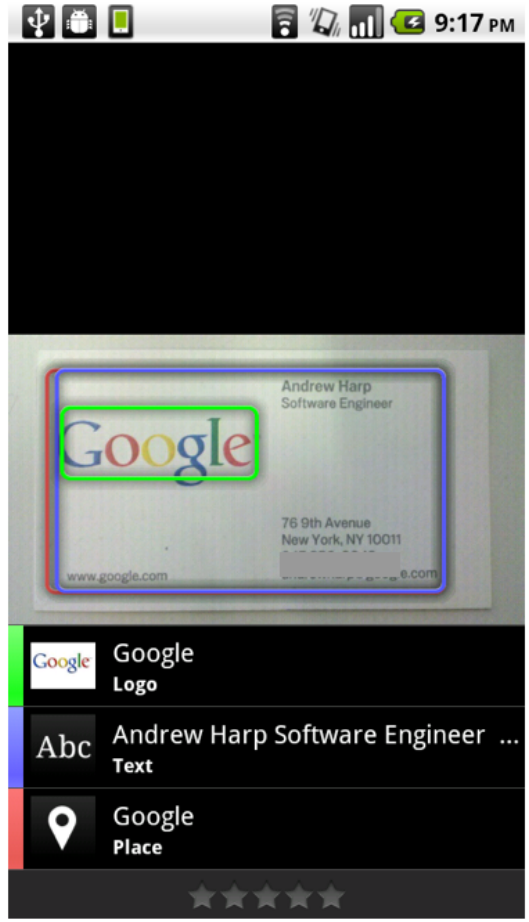
all results



Business card, streaming results



Text
Landmark
Similar Image
Logo
request roundtrip



If we had faster hardware...

- Textured 2D recognition improvements
  - More complex features used in recognition
  - 3D modeling to help constrain the search space

- OCR
  - Fewer timeouts mean greater accuracy and determinism
  - Larger dicts, multiple binarizers, overlapping patches
  - Matching with all possible characters at all fonts

- Improved tracking
  - Sudden / fast movements in low-textured regions
  - AR without known textured objects

- Selecting the best frame to process
  - Superresolution to enhance resolution / remove noise

Hardware asks, cont.

- Client-side
  - Fine flash control
  - FoV, whitebalance, depth estimation, ...
  - Reducing time-to-picture
    - Faster focus, Faster shutter and writing to memory
  - Faster GPUs / short video-memory read time

- Server-side
  - Memory bandwidth & larger caches
    - For feature extraction / classification
  - Greater parallelism: SIMD for basic linear algebra
  - Handle higher-resolution images

Potential future, given such advances

- Ease friction in financial transactions
  - Photograph checks, credit card numbers
- Use a front facing camera to detect if the user is happy
- Solve a Sudoku puzzle

- What would it mean if the camera was always on?
  - Replay that lecture I attended
  - Recall the name of the restaurant I ate during trip
  - Collaborative 3D model reconstruction
  - New twist on smart dust?

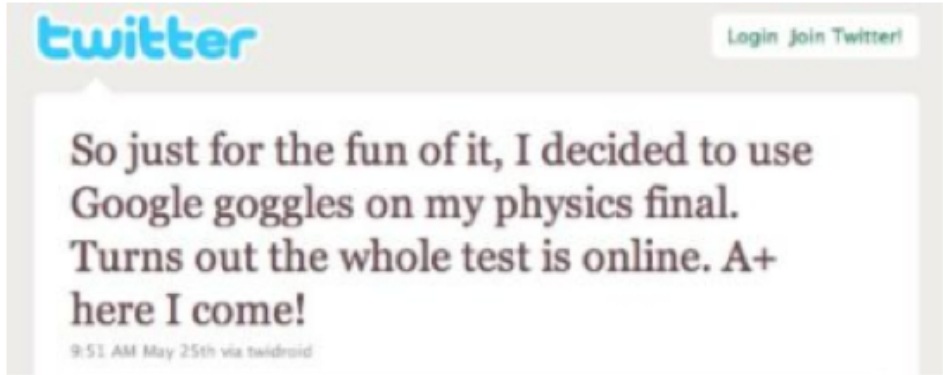
- Don't forget battery life

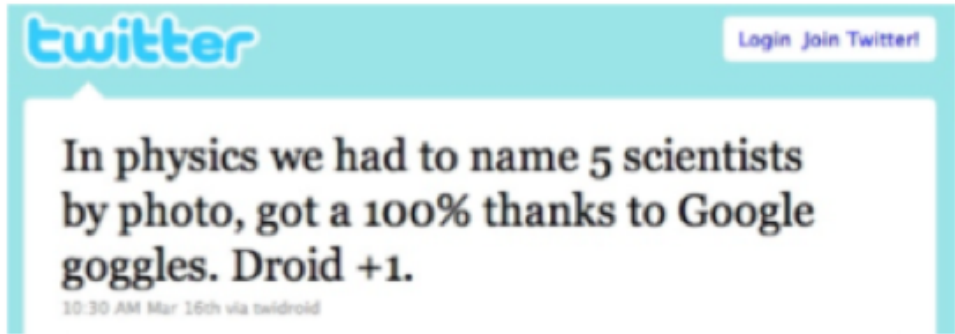
Super human abilities

- Sheer scale
  How many artworks do we know?
- High quality labels (sometimes)
- Added intelligence
  Translate
- Killer app
  Find online versions of tests
- *Soon there will be a companion looking over your shoulder knowing more about every item in your field of view!*

Super human abilities

- Sheer scale
  How many artworks do we know?
- High quality labels (sometimes)
- Added intelligence
  Translate
- Killer app
  Find online versions of tests
- *Soon there will be a companion looking over your shoulder knowing more about every item in your field of view!*





Goggles going mainstream

- Smartphone needed
  "Smartphone sales to end users accounted for 19 percent of worldwide mobile device sales, an increase of 50.5 percent from the second quarter of 2009."
- Cheap and fast network needed

Table 2
Worldwide Smartphone Sales to End Users by Operating System in 2Q10 (Thousands of Units)

Company	2Q10 Units	2Q10 Market Share (%)	2Q09 Units	2Q09 Market Share (%)
Symbian	25,386.8	41.2	20,880.8	51.0
Research In Motion	11,228.8	18.2	7,782.2	19.0
Android	10,606.1	17.2	755.9	1.8
iOS	8,743.0	14.2	5,325.0	13.0
Microsoft Windows Mobile	3,096.4	5.0	3,829.7	9.3
Linux	1,503.1	2.4	1,901.1	4.6
Other OSs	1,084.8	1.8	497.1	1.2
Total	**61,649.1**	**100.0**	**40,971.8**	**100.0**

Source: Gartner (August 2010)

Goggles availability, cont.

- iPhone version by end of year
- Significant cost to writing client applications
  - Separate code bases for each platform
  - Difficult to test -> infrequent version updates
- Is there a standard platform? It's called a browser.
- Can Goggles be a Web app?
  - Need fine control of camera. Some HTML5 support.
    - http://www.w3.org/TR/html-media-capture/
  - Need device sensors.
  - Need fine control of network.
  - Where should computation happen?

Roadmap

- Progress toward universal visual search
  - When it works, it's brilliant; but it does not always work
  - Increased coverage and accuracy (algo improvements)
  - Increase level of cross-engine inference
- Allow for 3rd party feeds and user annotations: self-service interface to add pics to Goggles
- Third-party APIs
  - Currency converter?
  - An app to go through photo album, Goggling each pic?
  - What would you use it for?
- Augmented reality presentation when it is the superior UI
- Combine external and internal searches
- Audio-visual search

Recap design principles

Universal

- Makes it more difficult compared to vertical solutions
- Needs very low false positive rate

To the degree possible, do not force the user to select modes

Specificity

- Object instance more important than object class recognition

Put best foot forward

Recall, precision, scale, and speed

Thank you!

Alessandro Bissacco, Anand Pillai, Andrew Harp, Andrew Hogue, Andrew Rabinovich, Anthony Sciola, Casey Ho, Chuck Rosenberg, David Petrou, Fernando Brucher, Gabe Taubman, Hartmut Neven, Hartwig Adam, Henry Rowley, Jiayong Zhang, Johannes Steffens, John Flynn, Laura Garcia-Barrio, Lijia Jin, Matt Bridges, Matt Casey, Max Braun, Mihai Badoiu, Rafael Spring, Sergey Ioffe, Shailesh Nalawadi, Ulrich Buddemeier, Xiaotao Duan, Yuan Li